Evaluating, Synthesizing, and Enhancing for Customer Support Conversation

Jie Zhu[1,2*], **Huaixia Dou**[1,2*], **Junhui Li**[1†], **Lifan Guo**[2],
Feng Chen[2], **Chi Zhang**[2], **Fang Kong**[1]

[1]School of Computer Science and Technology, Soochow University
[2]Qwen DianJin Team, Alibaba Cloud Computing
zhujie951121@gmail.com, lijunhui@suda.edu.cn

Abstract

Effective customer support requires not only accurate problem-solving but also structured and empathetic communication aligned with professional standards. However, existing dialogue datasets often lack strategic guidance, and real-world service data is difficult to access and annotate. To address this, we introduce the task of Customer Support Conversation (CSC), aimed at training customer service supporters to respond using well-defined support strategies. We propose a structured CSC framework grounded in COPC guidelines, defining five conversational stages and twelve strategies to guide high-quality interactions. Based on this, we construct `CSConv`, an evaluation dataset of 1,855 real-world customer–agent conversations rewritten using LLMs to reflect deliberate strategy use, and annotated accordingly. Additionally, we develop a role-playing approach that simulates strategy-rich conversations using LLM-powered roles aligned with the CSC framework, resulting in the training dataset `RoleCS`. Experiments show that fine-tuning strong LLMs on `RoleCS` significantly improves their ability to generate high-quality, strategy-aligned responses on `CSConv`. Human evaluations further confirm gains in problem resolution.

Code — https://github.com/aliyun/qwen-dianjin
Datasets — https://huggingface.co/DianJin
Extended version — https://arxiv.org/abs/2508.04423

1 Introduction

Customer support aims to help users resolve product- or service-related issues through effective and context-aware communication. While large-scale dialogue systems have received growing attention in recent years (Budzianowski et al. 2018; Rashkin et al. 2019; Peskov et al. 2019; Liu et al. 2021), customer support conversations remain underexplored in the NLP community, largely due to the scarcity of publicly available benchmarks and the sensitive, domain-specific nature of support interactions. Meanwhile, large language models (LLMs) like GPT-3 (Brown et al. 2020) have shown impressive capabilities in open-domain dialogue generation (Zhang et al. 2020b; Bae et al. 2022; Thoppilan et al. 2022), but their ability to generate realistic and effective customer support conversations remains underexamined.

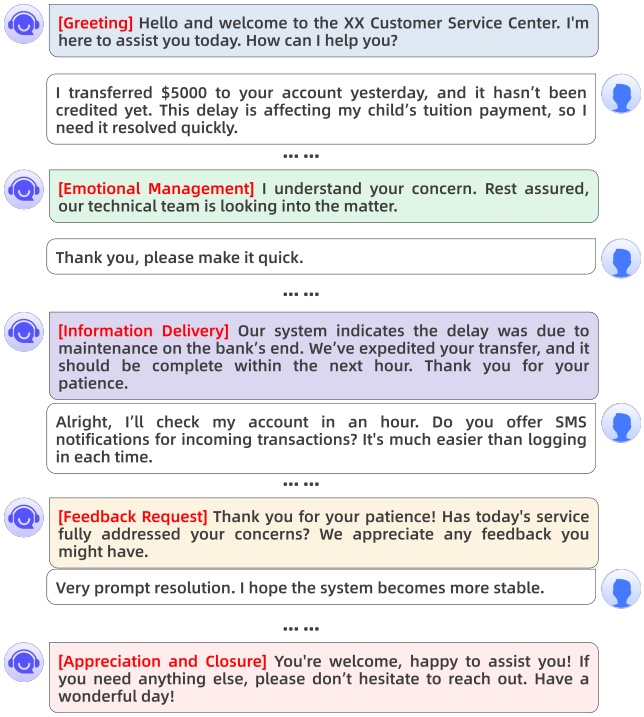

Figure 1: An example dialogue between a service supporter (left) and a customer (right), showing **support strategies** (noted in parentheses) used by the supporter. The conversation is organized into five stages of the proposed CSC framework, shown in colored boxes.

Customer support guidelines from COPC, the internationally recognized standard for customer experience management,[1] emphasize that high-quality support often depends on structured communication strategies, similar to those used in emotional support settings (Liu et al. 2021). As shown in Figure 1,[2] the service supporter begins with the *Greeting* strategy to initiate the conversation and explore the cus-

*These authors contributed equally.
†Corresponding Author.

[1]https://www.copc.com/copc-standards/
[2]The example is translated into English for better readability.

tomer's issue. Upon recognizing the customer's negative emotion, the supporter adopts the *Emotional Management* strategy, expressing empathy and understanding to alleviate the customer's distress. After understanding the problem, the supporter employs the *Information Delivery* strategy to provide clear and actionable guidance, followed by the *Feedback Request* strategy to check for further concerns. Finally, the conversation ends with the *Appreciation and Closure* strategy, ensuring a respectful and reassuring followed by *Feedback Request* to check for further concerns.

Despite the importance of effective customer support, research on real-time customer service conversations remains limited, mainly due to a lack of task-specific design and high-quality annotated data. Many studies (Xu et al. 2017; Oraby et al. 2017; Cui et al. 2017; Mesquita, Martins, and Almeida 2022) rely on asynchronous, exchanges (e.g., Twitter), where interactions span minutes to days. These settings differ significantly from the immediate, uninterrupted nature of real-time support. Moreover, effective support requires not only resolving issues but also showing empathy and emotional support. Yet, most task-oriented dialogue datasets (Wen et al. 2017; Budzianowski et al. 2018; Peskov et al. 2019; Gung et al. 2023) lack the intentional use of supportive strategies like emotional management or empathetic closure, which are vital for high-quality customer service.

To address this gap, we introduce the task of **c**ustomer **s**upport **c**onversation (CSC), to facilitate the training of customer service supporters. The goal is to help them respond with appropriate strategies that combine accurate solutions with empathetic communication. Based on COPC standards and inspired by the emotional support framework (Liu et al. 2021), we develop a CSC-specific framework with five dialogue stages and twelve support strategies. Using it, we build `CSConv`, a high-quality dataset adapted from real service interactions and refined for structured strategy use. To address the lack of high-quality training data, we also develop a role-playing approach that creates `RoleCS`, a synthetic dataset of strategy-rich conversations by assigning LLMs distinct roles aligned with the CSC framework. Fine-tuning on `RoleCS` significantly boosts LLMs' ability to generate strategy-aligned and effective responses on `CSConv`.

2 Related Work

2.1 Task-Oriented Conversation Datasets

The goal of `CSConv`, which is to generate conversations based on real spoken customer service interactions, differs significantly from previous task-oriented dialogue datasets. Many earlier synthetic datasets have been created using the Wizard of Oz (WOZ) framework (Kelley 1984), where one person acts as the system and another as the user. Wen et al. (2017) introduce a crowdsourced version of the WOZ framework to collect domain-specific dialogue data more efficiently. Other task-oriented dialogue datasets include Frames (El Asri et al. 2017), MultiWOZ (Budzianowski et al. 2018), MultiDoGO (Peskov et al. 2019), EMPATHETICDIALOGUES (Rashkin et al. 2019), Taskmaster-1 (Byrne et al. 2019), ESConv (Liu et al. 2021), and NATCS (Gung et al. 2023), each of which explores different domains and annotation schemes for developing conversational agents.

Stage	Description
Connecting	Greeting and establishing connection
Identifying	Understanding and identifying problems
Exploring	Seeking solution
Resolving	Resolving and confirming
Maintaining	Ending and maintaining relationship

Table 1: Five stages in the CSC framework.

Among them, ESConv is closely related to our work, as both it and `CSConv` aim to enhance service quality using structured support strategies. However, their construction methods differ significantly: ESConv uses a crowdsourced WOZ setup, while `CSConv` rewrites real-world service dialogues using high-performing LLMs.

2.2 Role-Playing in Conversation Generation

Recent advances in LLMs have enabled the use of role-playing agents for conversation generation. For example, Bae et al. (2022) generate dialogues aligned with specific roles, while Yang et al. (2024b) simulate characters and settings to produce coherent interactions. Wu et al. (2024) create interactive narrative dramas, and Ye et al. (2025) focus on emotional support conversations. Building on the Ye et al. (2025), we design role-playing agents for customer support with clearer role definitions and responsibilities to more closely reflect real-world service interactions.

3 CSC: Customer Support Conversation

In customer support conversation (CSC), the service supporter aims to accurately identify customer issues while also addressing emotional needs. This dual focus enables effective solutions that enhance customer satisfaction and improve problem-resolution efficiency. The supporter is considered effective when both the issue and emotional concerns are properly addressed. To build an evaluation dataset for CSC, we first introduce the CSC framework in Section 3.1, followed by the data construction process in Section 3.2 and dataset statistics in Section 3.3. Finally, we define the CSC task in Section 3.4.

3.1 CSC Framework

The CSC framework organizes the customer support process into five stages, each with recommended strategies designed to enhance the quality and effectiveness of interaction.

Stages. Building on the three core stages of supportive communication from Hill (2019) and the COPC practical guidelines, we work with domain experts to refine and extend this structure for customer support. As shown in Table 1, the resulting CSC framework defines five stages: *Connecting* (greeting and rapport building), *Identifying* (understanding the customer's issue and emotional state), *Exploring* (discussing and evaluating potential solutions), *Resolving* (delivering and confirming resolution), and *Maintaining* (closing the interaction while preserving the customer relationship). Figure 2 illustrates the flow across these stages.

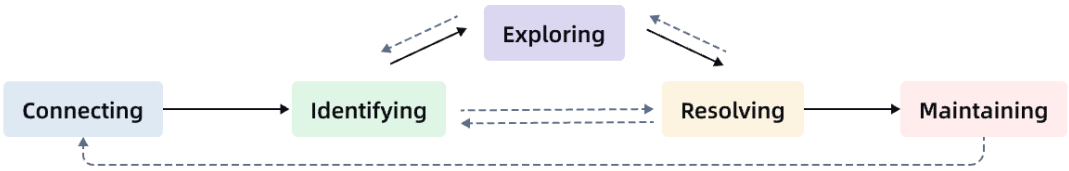

Figure 2: Overview of the CSC framework's five stages, each paired with recommended support strategies (see Table 2). The typical flow is: ① Connecting → ② Identifying → ③ Exploring → ④ Resolving → ⑤ Maintaining (black arrows), but it can be adjusted based on the specifics of each conversation (dashed arrows).

Strategy	Stages	Description
Greeting (GT)		Utilize friendly and professional language to greet customers, creating a warm communication atmosphere.
Identity Verification (IV)		Ensure the accuracy and security of the service by asking for the customer's basic information.
Emotional Management (EM)		Express understanding and care for the customer's feelings to help alleviate negative emotions.
Restatement or Paraphrasing (RP)		Restate the customer's issue to ensure accurate understanding.
Problem Refinement (PR)		Employ detailed inquiries to fully and accurately comprehend customer needs.
Providing Suggestions (PS)		Offer professional advice or action steps based on the customer's issue.
Information Delivery (ID)		Clearly explain relevant company policies, processes, or steps to help customers understand the basis of solutions.
Resolution Implementation (RI)		Execute the agreed-upon solution, ensuring all steps are followed as planned, and update the customer on the progress.
Feedback Request (FR)		Seek customer feedback after the issue has been addressed to gauge their satisfaction and identify potential areas for improvement.
Appreciation and Closure (AC)		End the conversation positively, ensuring the customer feels valued and laying a solid foundation for future interactions.
Relationship Continuation (RC)		Guide customers towards future service or product updates, ensuring they understand how to continue receiving support and service in the future, thereby establishing a bridge for further interaction.
Others		Situations that do not belong to the above eleven strategies.

Table 2: Twelve strategies in the CSC framework. The cells of lightblue, lightgreen, lightpurple, lightyellow, and lightpink represent the Connecting, Identifying, Exploring, Resolving, and Maintaining stages, repsectively.

Importantly, these stages are not rigid steps but modular components that can appear in various orders or combinations depending on the nature of the conversation. For example, even when no solution is reached, a supporter may still provide empathy (*Exploring*), acknowledge service limitations (*Resolving*), and close the conversation professionally (*Maintaining*). This flexible structure enables consistent analysis and strategy modeling across both successful and challenging customer support scenarios.

Strategies. In parallel, we work with domain experts to define twelve actionable support strategies aligned with the five stages: *Greeting (GT)*, *Identity Verification (IV)*, *Emotional Management (EM)*, *Restatement or Paraphrasing (RP)*, *Problem Refinement (PR)*, *Providing Suggestions (PS)*, *Information Delivery (ID)*, *Resolution Implementation (RI)*, *Feedback Request (FR)*, *Appreciation and Closure (AC)*, *Relationship Continuation (RC)*, and *Others*. As shown in Table 2, these strategies provide practical guidance for managing both task-related issues and emotional support throughout the customer interaction.

3.2 Dataset Construction

We construct the `CSConv` dataset using Chinese customer service dialogues collected from our pre-sales and after-sales customer service centers. Prior to our access, all conversations were professionally transcribed, manually corrected, and fully de-identified to ensure privacy protection and integrity. While these raw conversations authentically reflect real-world customer interactions, they often lack consistent structure, making it difficult to systematically annotate support strategies according to the COPC-informed CSC framework. To address this, we employ an LLM to rewrite the conversations while preserving the original semantics and user intent. This controlled rewriting aligns con-

versations with the defined dialogue stages and strategies, enabling more consistent and interpretable annotations without sacrificing the complexity of real user queries.

Importantly, the goal of constructing `CSConv` *is not to evaluate real-time chatbot performance, but to facilitate the training of customer service supporters by helping them learn to respond using appropriate strategies guided by the COPC framework.* In total, we collect 690K conversations spanning eight in-domain topics. To ensure quality, we use a four-stage pipeline to guide data selection and refinement.

1. Pre-filtering: We first apply rule-based filtering to remove low-quality conversations. These rules exclude conversations that are too short or too long, contain overly lengthy utterances, show an imbalance between customer and agent turns, or have a high proportion of ineffective customer responses. Additionally, we use an LLM to exclude conversations with offensive or unprofessional content.
2. Sampling and Rewriting: Up to 500 filtered conversations per topic are sampled and rewritten by an LLM to align with the CSC framework. During this process, the LLM analyzes the original scenario and generates a new conversation that preserves the core issue while improving clarity, structure, and emotional engagement. For each agent turn, the LLM selects an appropriate support strategy based on conversation context, occasionally using *Others* to maintain conversational naturalness. Customer responses are also refined for coherent interaction.
3. Post-filtering: After rewriting, a second round of rule-based and LLM-based checks verifies structure (e.g., strategy coverage, speaker alternation, and length) and filters out conversations lacking coherence, empathy, or strategic alignment.
4. Manually annotation: Finally, experts certified in COPC review the remaining conversations, evaluating the support agent's responses for realism, empathy, and adherence to the CSC framework. This results in a curated evaluation set of 1,855 high-quality conversations.

The `Technical Appendix` provides details on filtering rules, prompt templates, and annotation guidelines. For rewriting, we use `DeepSeek-R1` (Guo et al. 2025) instead of `GPT-4o` (OpenAI 2024), as the latter tends to produce shorter, less emotionally rich dialogues.

3.3 Dataset Analysis

Overview Statistics. Table 3 compares statistics of `CSConv` before and after rewriting, covering 1,855 conversations. Rewritten dialogues are longer, averaging 27.27 versus 19.06 utterances, with supporter responses increasing from 41.16 to 48.72 words and customer responses decreasing from 21.60 to 17.17. Notably, strategy use (excluding *Others*) rises from 55.28% to 97.82%,[3] indicating more deliberate and guided support. These shifts reflect the goal of rewriting, which is to facilitate the training of customer service supporters in applying appropriate strategies aligned with the COPC framework.

[3]We use `Qwen2.5-72B-Instruct` to assign a support strategy to each supporter response in the dialogues prior to rewriting.

-	-	Number	
		Original	**Rewritten**
Total	Conversations	1,855	1,855
	Utterances	35,350	50,587
	Avg. Utterance Number	19.06	27.27
	Avg. Utterance Length	31.48	33.27
Supporter	Utterances	17,862	25,810
	Avg. Utterance Number	9.63	13.91
	Avg. Utterance Length	41.16	48.72
	Strategy Use (w/o *Others*)	55.28%	97.82%
Customer	Utterances	17,488	24,777
	Avg. Utterance Number	9.43	13.36
	Avg. Utterance Length	21.60	17.17

Table 3: Statistics of `CSConv` before and after rewriting.

Topic	**Num**	**Proportion**
Account and Transaction Management	265	14.3%
Product Consultation	242	13.0%
Technical Support and Online Services	295	15.9%
Complaints and Dispute Resolution	263	14.2%
Marketing and Promotion Activities	211	11.4%
Risk Management and Security	266	14.3%
Financial Consulting and Planning	263	14.2%
Others	50	2.7%
Overall	1,855	100.0%

Table 4: Distribution of topics in `CSConv`.

Topic Distribution. Table 4 presents the distribution of conversations across eight topic. Each topic, excluding *Others*, accounts for roughly 11% to 16% of the dataset.

Strategy Distribution. As shown in Figure 3, the most commonly used strategies are *Information Delivery* (14.9%), *Emotional Management* (11.9%), and *Provide Suggestions* (10.0%). This highlights the dual importance of delivering clear information and managing customer emotions in effective support conversations.

Additional statistics on strategy transitions and unique strategy distributions are provided in the `Technical Appendix`.

3.4 CSC Task Definition

We denote a customer support conversation as $D = \{(P_i, T_i, U_i)\}_{i=1}^{N}$, consisting of N turns exchanged between a supporter S and a customer C. For each turn, $P_i \in \{S, C\}$ denotes the speaker, U_i is the utterance text, and T_i represents the response strategy used. Strategies are selected from a predefined set G, and are assigned only to the supporter's turns. That is, if $P_i = C$, then $T_i = \text{NULL}$.

Following the task of emotional support conversation (Liu et al. 2021; Ye et al. 2025), we define the CSC task as generating the supporter's response. Specifically, at turn k, where $P_k = S$, the model receives the conversation history $X_k = \{(P_i, T_i, U_i)\}_{i=1}^{k-1}$ as input. The CSC task is then divided into two subtasks:[4]

[4]We further investigate how strategy prediction contributes to improving response generation in Section 6.3.

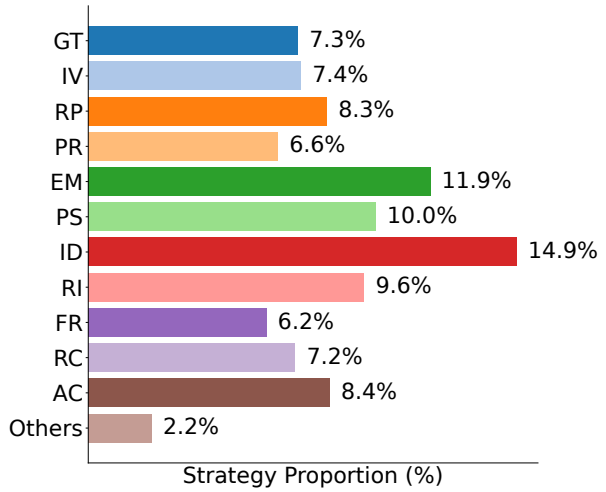

Figure 3: Strategy proportion of CSConv.

1. **Strategy Prediction:** Predict the appropriate support strategy $T_k \in G$ based on the conversation history X_k.
2. **Response Generation:** Generate a response U_k conditioned on both the predicted strategy T_k and the conversation history X_k, ensuring that the output is consistent with both the customer's needs and the strategic intent.

4 Synthetic Conversation Generation with Role-Playing Agents

Fine-tuning LLMs has shown effective for improving performance in task-specific dialogue applications. However, building high-quality, multi-turn customer support datasets remains challenging due to the need for domain expertise, contextual coherence, and dialogue diversity. This limits both the scalability and scenario coverage. While recent studies use LLMs for dataset augmentation via rewriting or imitation, they often yield limited variation in dialogue flow and support strategy. To address this, inspired by Ye et al. (2025), we adopt a multi-role role-playing framework to generate synthetic customer support dialogues that are diverse, coherent, and representative of real-world scenarios.

4.1 Role Playing Conversation Generation

As shown in Figure 4, our role-playing framework involves five roles: *Planner*, *Supporter Assistant*, *Supporter*, *Customer Assistant*, and *Customer*. Each role serves a distinct purpose. The *Planner* defines the dialogue scenario and sets the *Customer*'s communication goal. The main interaction occurs between the *Supporter* and the *Customer*, simulating a real customer service exchange. Meanwhile, their respective *Assistants*, the *Supporter Assistant* and the *Customer Assistant*, offer strategic guidance to help them fulfill their roles more effectively. We use `deepseek-r1` (Guo et al. 2025) to simulate all roles, with detailed prompts available in the `Technical Appendix`. We describe each role below.

Planner Before the conversation begins, the Planner selects a topic e from a predefined list E of customer topics in Table 4 (excluding *Others*). It then samples a customer profile o from a character pool O that contains diverse customer personas. Using the selected topic e and customer profile o, the Planner prompts an LLM $\mathcal{M}$ to generate a detailed service scenario e' along with a corresponding communication goal g, denoted as $(g, e') = \mathcal{M}(o, e)$. This setup ensures that each dialogue is grounded in a realistic context.

Supporter Assistant The Supporter Assistant recommends an appropriate support strategy t from the predefined strategy set G, based on the current dialogue history h_s and the scenario e'. The recommendation is generated by querying the LLM $\mathcal{M}$, i.e., $t = \mathcal{M}(h_s, G, e')$. This strategic guidance helps the Supporter stay aligned with the customer's needs and maintain coherent throughout the conversation.

Supporter Guided by the support strategy t suggested by the Supporter Assistant, the Supporter generates a contextually appropriate response r_s. This response is produced by the LLM $\mathcal{M}$, conditioned on the dialogue history h_s, the strategy t and the scenario e', i.e., $r_s = \mathcal{M}(h_s, t, e')$.

Customer Assistant The Customer Assistant guides the conversation by generating the next direction d, ensuring alignment with the customer's communication goal g. To do so, it prompts the LLM $\mathcal{M}$ using the current dialogue history h_c, the customer's goal g, and the scenario e'. This helps the customer maintain coherent and goal-oriented behavior throughout the conversation. Formally, the direction is generated as $d = \mathcal{M}(h_c, g, e')$.

Customer The Customer generates a response r_c based on the current dialogue history h_c, the intended conversational direction d, the character profile o, and the scenario e'. This ensures that the response is both contextually appropriate and consistent with the customer's persona. Formally, the generation process is denoted as: $r_c = \mathcal{M}(h_c, d, o, e')$.

Diversity in conversations relies heavily on the richness of customer personas (Wang et al. 2025). To support this, we construct a rich character profile pool O that captures a wide spectrum of customer backgrounds, behaviors, and communication styles, thereby enhancing the realism and variability of the simulated interactions.

Character Profile Pool. We design a comprehensive character profile template specifically for customer personas, incorporating attributes such as demographics, financial status, and communication preference. To populate this template, we use `Qwen2.5-72B-Instruct` to automatically extract and complete profile information from 15,980 real-world customer service dialogues. To reduce redundancy and ensure profile diversity, each structured profile is converted into a free-text description, and pairwise cosine similarity is computed using Qwen's `text-embedding-v2`. Profiles exceeding a similarity threshold of 0.85 are considered redundant and removed. Following this filtering process, we retain 1,948 distinct customer profiles in our final pool. Additional details, including the profile template

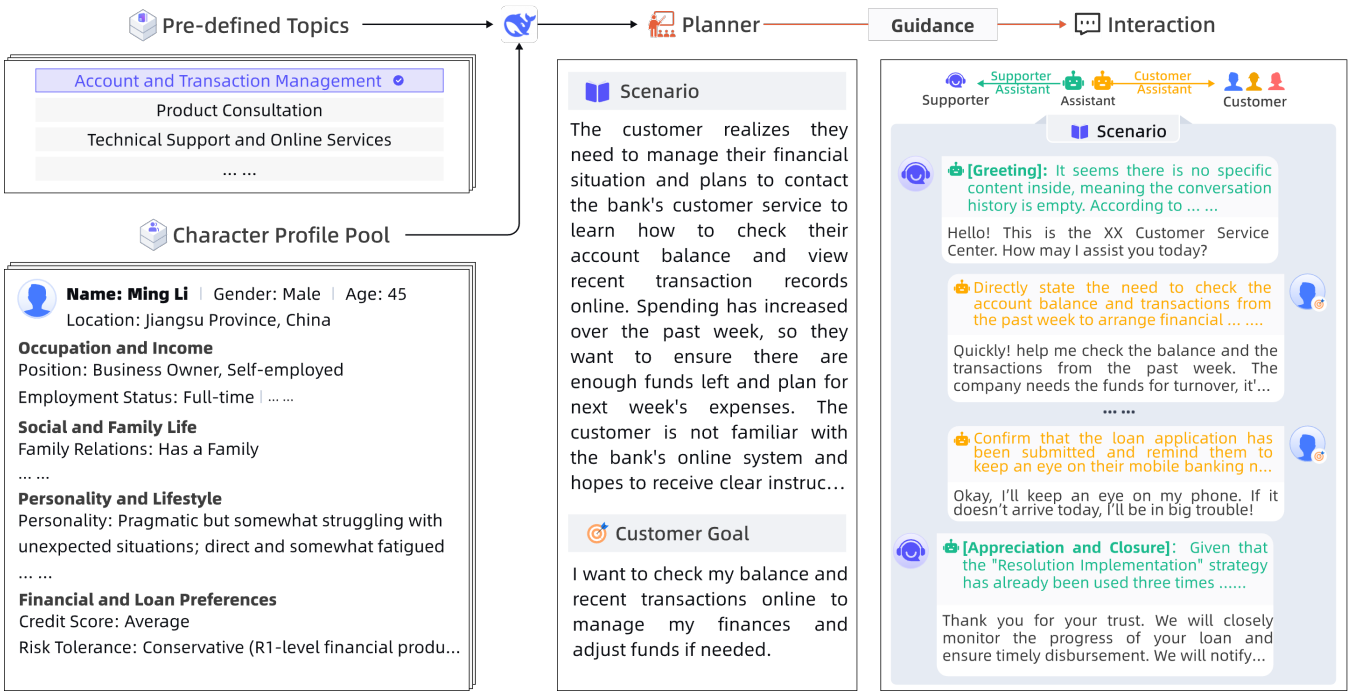

Figure 4: Illustration of synthetic conversation generation using role-playing agents.

-	-	Number
Total	Dialogues	11,232
	Utterances	263,580
	Avg. Utterance Number	23.47
	Avg. Utterance Length	57.14
Supporter	Utterances	137,406
	Avg. Utterance Number	12.23
	Avg. Utterance Length	66.98
Customer	Utterances	126,174
	Avg. Utterance Number	11.23
	Avg. Utterance Length	46.43

Table 5: Statistics of the RoleCS dataset.

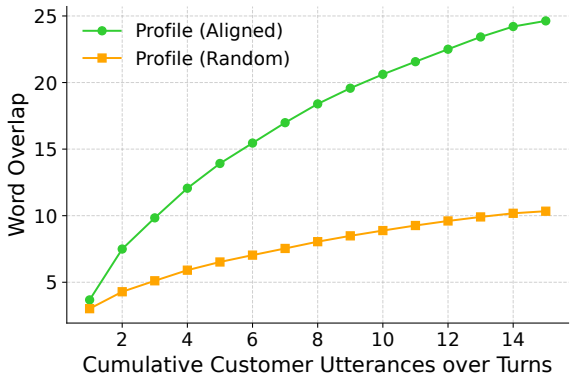

Figure 5: Word overlap between cumulative customer utterances and profile (Aligned vs. Random).

and construction prompts, are provided in the Technical Appendix.

4.2 Synthetic Dataset

The Planner generates one conversation per unique pair of customer topic and profile. With $(|E|-1)$ topics (excluding *Others*) and $|O|$ profiles, this yields 13,636 conversations. After applying quality filters detailed in the Technical Appendix, we retain 11,232 high-quality conversations, forming the final training (or fine-tuning) dataset, RoleCS. Table 5 shows the statistics of RoleCS. To assess whether customer utterances reflect assigned profiles, we randomly sample 500 dialogues and measure word overlap between customer utterances and their corresponding aligned profiles, using random profiles for comparison. As shown in Figure 5, word overlap with aligned profiles rises rapidly at first as profile information is elicited, then increases more gradually once most information has been mentioned. Throughout, word overlap with aligned profiles remains consistently higher than with random profiles, indicating that customer utterances naturally reflect profile details. See the Technical Appendix for more on RoleCS.

5 Experimentation

Test-Time Prompting. For both subtasks outlined in Section 3.4, we adopt a unified single-prompt method. Given the conversation history, the LLM, whether fine-tuned or not, is prompted once to first identify the appropriate support strategy and then generate the corresponding response. The full prompt is provided in the Technical Appendix.

Model	Size	B-2	B-4	R-L	BS	BR	ACC	B-2	B-4	R-L	BS	BR	ACC
		Evaluation with reference context						*Evaluation with generated context*					
GPT-4o	-	8.13	2.97	4.12	64.26	51.27	42.58	6.41	2.07	2.22	62.95	51.48	36.29
DeepSeek-R1	671B	11.67	5.09	**8.44**	66.57	52.09	39.78	8.41	3.11	**6.27**	64.80	52.01	35.23
DeepSeek-V3	671B	11.57	4.95	7.04	66.09	53.10	41.99	8.43	3.33	4.14	64.13	53.09	36.54
LLaMA3.1-Instruct	8B	4.28	1.44	3.62	58.68	38.84	17.16	2.62	0.89	1.31	55.06	35.97	13.75
+ RoleCS	8B	11.06	4.77	6.93	66.21	52.13	42.15	8.80	3.72	4.52	64.57	49.85	35.73
LLaMA3.1-Instruct	70B	6.85	2.38	4.10	63.14	47.53	38.78	5.57	1.76	2.59	62.24	46.83	30.36
+ RoleCS	70B	11.73	5.11	7.64	66.62	53.69	42.79	**9.62**	**4.00**	4.89	**65.15**	50.91	**39.44**
Qwen2.5-Instruct	7B	6.73	2.29	3.80	62.75	48.21	18.78	5.11	1.53	2.43	61.62	47.43	19.39
+ RoleCS	7B	11.23	4.80	7.39	66.35	52.14	42.96	8.61	3.51	4.13	64.42	48.80	30.52
Qwen2.5-Instruct	72B	8.61	3.23	5.41	64.63	52.49	37.22	6.28	2.00	3.59	63.14	50.55	30.93
+ RoleCS	72B	**12.15**	**5.32**	7.97	**66.85**	**54.49**	**43.29**	9.38	3.90	4.35	64.89	**53.65**	36.02

Table 6: Performance comparison on CSConv. The best and second best results are in **bold** and underlined, respectively.

5.1 Experimental Settings

Models. We evaluate several widely used LLMs on the CSC task, including GPT-4o (OpenAI 2024), DeepSeek-R1 (Guo et al. 2025), DeepSeek-V3 (Liu et al. 2024), Qwen-2.5-7B/72B-Instruct (Yang et al. 2024a), and LLaMA-3.1-8B/70B-Instruct (Grattafiori et al. 2024). To evaluate the impact of the proposed RoleCS, we fine-tune the Qwen and LLaMA models using 137,406 fine-tuning instances extracted from RoleCS, formatted consistently with our single-prompt in test-time. Fine-tuning and inference configurations are provided in the Technical Appendix.

Metrics. Following prior work (Liu et al. 2021; Ye et al. 2025), we adopt a diverse set of metrics to assess the response quality. These include BLEU-n (**B-n**) (Papineni et al. 2002) and ROUGE-L (**R-L**) (Lin 2004) for measuring lexical overlap, BERTScore (**BS**) (Zhang et al. 2020a) and BLEURT (**BR**) (Sellam, Das, and Parikh 2020) for semantic similarity. In addition, to evaluate alignment with the intended support strategy, which is a core aspect of the CSC task, we report strategy prediction accuracy (**ACC**).

5.2 Experimental Results

Table 6 presents the main results on CSConv under two evaluation settings: (1) ***evaluation with reference context***, which all LLMs are evaluated using the same gold history for fair comparison, and (2) ***evaluation with generated context***, which assesses performance when relying on model-generated histories, thus reflecting the ability to maintain coherence and relevance without ground truth context. From the results, we have the following observations:

- Larger models tend to perform better among non–fine-tuned LLMs. Additionally, Chinese-centric models like Qwen and DeepSeek outperform more general models such as LLaMA and GPT, indicating that alignment with language and cultural context benefits CSC performance.
- RoleCS proves highly effective, as fine-tuning on it significantly improve performance across all metrics. In particular, Qwen2.5-Instruct-72B, after fine-tuning, matches or surpasses DeepSeek-R1, a strong baseline for Chinese-language tasks.
- Evaluation with generated context shows similar performance trends as with reference context, but with lower absolute scores. This drop reflects the difficulty of maintaining consistency and quality in multi-turn conversations when relying on model-generated dialogue history, echoing findings in prior work such as Ye et al. (2025).

SFT Data	B-2	B-4	R-L	BS	BR	ACC
Baseline 1	8.51	3.09	5.35	64.35	46.70	33.09
Baseline 2	9.40	3.26	6.98	65.22	50.18	36.17
RoleCS	**11.23**	**4.80**	**7.39**	**66.35**	**52.14**	**42.96**

Table 7: Performance comparison of fine-tuning on different datasets.

6 Discussion

We conduct further analysis under the reference context setting to examine the effectiveness of our role-playing approach in synthetic conversation generation.

6.1 Impact of Role-Playing on Data Quality

To evaluate our role-playing framework, we compare RoleCS with two baselines:

- Baseline 1: Conversations are generated via in-context learning without any role-playing.
- Baseline 2: Role-playing is applied, but without the Supporter Assistant agent.

All three datasets are generated using Deepseek-R1 with equal conversation counts. For fairness, we fine-tune Qwen2.5-7b-Instruct on each. As shown in Table 7, role-playing improves performance over in-context learning (Baseline 2 vs. Baseline 1), and adding the Supporter Assistant further boosts results (RoleCS vs. Baseline 2). These results highlight the value of strategy-rich training data and the importance of high-quality dataset construction.

6.2 Effect of Synthetic Dataset Size

To understand how dataset size influences performance, we split the RoleCS dataset into subsets of {0, 3K, 6K, 9K,

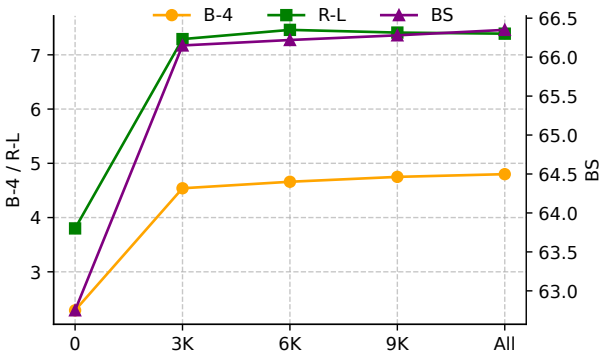

Figure 6: Performance comparison under different synthetic conversation dataset sizes.

Strategy	B-4	BS	ACC
Vanilla	3.11	64.43	-
Predict	3.23	64.63	37.22
Oracle	**3.80**	**66.34**	**100.00**

Table 8: Performance comparison across different support strategy variants.

All} dialogues. We fine-tune `Qwen2.5-7B-Instruct` on each subset individually. As shown in Figure 6, the most significant gains occur with the initial 3K examples. Beyond this point, performance improvements become marginal, particularly for ROUGE-L. These results suggest that even a modest amount of high-quality synthetic data can bring strong gains, with limited benefits from adding more.

6.3 Impact of Support Strategy Guidance

We evaluate the effect of incorporating support strategies using three variants of `Qwen2.5-72B-Instruct` without fine-tuning: (1) Vanilla, where the model generates a response without any support strategy guidance; (2) Predict, our default setup, where the model first predicts the strategy and then generates a response; and (3) Oracle, where the ground-truth strategy is provided in the prompt.

As shown in Table 8, the Predict variant slightly outperforms the Vanilla, indicating that even simple strategy prediction improves response quality. The Oracle variant performs best, highlighting that more accurate strategy prediction can further enhance the performance of CSC task.

6.4 Evaluation with LLMs and Human as Judges

To mitigate potential bias (e.g., GPT-4o favoring its own outputs), we use both GPT-4o and Qwen-Plus[5] to evaluate response quality across six dimensions: accuracy, helpfulness, understanding, coherence, informativeness, and empathy, each scored on a 0–100 scale. We report the average overall scores from both GPT-4o-Judge and Qwen-Plus-Judge, using the evaluation prompt in the `Technical Appendix`. As shown in Table 9, both judges reveal consistent performance patterns. Notably, fine-tuned `Qwen2.5-Instruct-72B` and `LLaMA3.1-Instruct-70B` outperform other models, even exceeding `DeepSeek-R1` in overall quality.

[5]https://help.aliyun.com/zh/model-studio/models

Model	Size	GPT-4o	Qwen-Plus	Human
DeepSeek-R1	671B	90.89	89.70	3.55
DeepSeek-V3	671B	90.54	89.54	3.36
LLaMA3.1-InstructΔ	8B	90.34	88.68	2.93
LLaMA3.1-InstructΔ	70B	91.02	89.76	3.58
Qwen2.5-InstructΔ	7B	90.46	88.76	3.10
Qwen2.5-InstructΔ	72B	**91.04**	**89.98**	**3.79**

Table 9: Model performance evaluated by GPT-4o, Qwen-Plus, and human judges. Δ denotes fine-tuned models.

Agreement	Kappa
GPT-4o-Judge & Qwen-Plus	0.726
GPT-4o-Judge & Human Annotators	0.658
Human Annotators	0.628

Table 10: Fleiss' Kappa scores.

For human evaluation, we randomly select 100 conversations and have them independently rated by three professional annotators on a 1-5 Likert scale (Joshi et al. 2015) across the same six dimensions. Table 9 shows the average overall scores, which align with trends from the LLM-based evaluations. Table 10 reports Fleiss' Kappa (Fleiss 1971) scores among annotators and between `GPT-4o-Judge` and the annotators. The results show strong inter-rater agreement and confirm that `GPT-4o-Judge` aligns well with both Qwen-Plus and human judgment.

7 Conclusion

This paper addresses the challenges of customer support conversations (CSCs) in the NLP field by introducing `CSConv`, a high-quality dataset grounded in support strategies, and a role-playing framework for generating realistic, goal-driven dialogues. Our approach significantly enhances LLMs' ability to generate coherent, context-aware, and empathetic responses in customer service scenarios. We believe that our contributions will inspire further research in this field, promoting advancements in empathetic and effective customer support technologies.

Acknowledgments

The authors express their heartfelt gratitude to the anonymous reviewers. We also offer our sincere appreciation to the members of the Qwen DianJin Team for their exceptional contributions, dedication, and hard work, which were instrumental in the success of this project. This work was supported by the Alibaba Innovative Research Program, and the National Natural Science Foundation of China (Grant No. 62276178), and the Key Project 23KJA520012 under the Natural Science Foundation of Jiangsu Higher Education Institutions.

References

Bae, S.; Kwak, D.; Kim, S.; Ham, D.; Kang, S.; Lee, S.-W.; and Park, W. 2022. Building a Role Specified Open-Domain Dialogue System Leveraging Large-Scale Language Models. In *Proceedings of NAACL-HLT*, 2128–2150.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language models are few-shot learners. In *Proceedings of NeurIPS*, 1877–1901.

Budzianowski, P.; Wen, T.-H.; Tseng, B.-H.; Casanueva, I.; Ultes, S.; Ramadan, O.; and Gašić, M. 2018. MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In *Proceedings of EMNLP*, 5016–5026.

Byrne, B.; Krishnamoorthi, K.; Sankar, C.; Neelakantan, A.; Goodrich, B.; Duckworth, D.; Yavuz, S.; Dubey, A.; Kim, K.-Y.; and Cedilnik, A. 2019. Taskmaster-1: Toward a Realistic and Diverse Dialog Dataset. In *Proceedings of EMNLP-IJCNLP*, 4516–4525.

Cui, L.; Huang, S.; Wei, F.; Tan, C.; Duan, C.; and Zhou, M. 2017. SuperAgent: A Customer Service Chatbot for E-commerce Websites. In *Proceedings of ACL, System Demonstrations*, 97–102.

El Asri, L.; Schulz, H.; Sharma, S.; Zumer, J.; Harris, J.; Fine, E.; Mehrotra, R.; and Suleman, K. 2017. Frames: a corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, 207–219.

Fleiss, J. L. 1971. Measuring Nominal Scale Agreement among Many Raters. *Psychological Bulletin*, 76(5): 378–382.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; and Others. 2024. The Llama 3 Herd of Models. *CoRR*, abs/2407.21783.

Gung, J.; Moeng, E.; Rose, W.; Gupta, A.; Zhang, Y.; and Mansour, S. 2023. NatCS: Eliciting Natural Customer Support Dialogues. In *Findings of ACL*, 9652–9677.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; et al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *CoRR*, abs/2501.12948.

Hill, C. E. 2019. *Helping Skills: Facilitating Exploration, Insight, and Action*. American Psychological Association, 5th edition.

Joshi, A.; Kale, S.; Chandel, S.; and Pal, D. K. 2015. Likert scale: Explored and explained. *British journal of applied science & technology*, 7(4): 396.

Kelley, J. F. 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Trans. Inf. Syst.*, 2(1): 26–41.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; et al. 2024. DeepSeek-V3 Technical Report. *CoRR*, abs/2412.19437.

Liu, S.; Zheng, C.; Demasi, O.; Sabour, S.; Li, Y.; Yu, Z.; Jiang, Y.; and Huang, M. 2021. Towards Emotional Support Dialog Systems. In *Proceedings of ACL-IJCNLP*, 3469–3483.

Mesquita, T.; Martins, B.; and Almeida, M. 2022. Dense Template Retrieval for Customer Support. In *Proceedings of COLING*, 1106–1115.

OpenAI. 2024. GPT-4o System Card. *CoRR*, abs/2410.21276.

Oraby, S.; Gundecha, P.; Mahmud, J.; Bhuiyan, M.; and Akkiraju, R. 2017. "How May I Help You?": Modeling Twitter Customer ServiceConversations Using Fine-Grained Dialogue Acts. In *Proceedings of IUI*, 343–355.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL*, 311–318.

Peskov, D.; Clarke, N.; Krone, J.; Fodor, B.; Zhang, Y.; Youssef, A.; and Diab, M. 2019. Multi-Domain Goal-Oriented Dialogues (MultiDoGO): Strategies toward Curating and Annotating Large Scale Dialogue Data. In *Proceedings of EMNLP-IJCNLP*, 4526–4536.

Rashkin, H.; Smith, E. M.; Li, M.; and Boureau, Y.-L. 2019. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In *Proceedings of ACL*, 5370–5381.

Sellam, T.; Das, D.; and Parikh, A. 2020. BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of ACL*, 7881–7892.

Thoppilan, R.; Freitas, D. D.; Hall, J.; Shazeer, N.; Kulshreshtha, A.; and Others. 2022. LaMDA: Language Models for Dialog Applications. *CoRR*, abs/2201.08239.

Wang, X.; Zhang, H.; Ge, T.; Yu, W.; Yu, D.; and Yu, D. 2025. OpenCharacter: Training Customizable Role-Playing LLMs with Large-Scale Synthetic Personas. *CoRR*, abs/2501.15427.

Wen, T.-H.; Vandyke, D.; Mrkšić, N.; Gašić, M.; Rojas-Barahona, L. M.; Su, P.-H.; Ultes, S.; and Young, S. 2017. A Network-based End-to-End Trainable Task-oriented Dialogue System. In *Proceedings of EACL*, 438–449.

Wu, W.; Wu, H.; Jiang, L.; Liu, X.; Zhao, H.; and Zhang, M. 2024. From Role-Play to Drama-Interaction: An LLM Solution. In *Findings of ACL*, 3271–3290.

Xu, A.; Liu, Z.; Guo, Y.; Sinha, V.; and Akkiraju, R. 2017. A New Chatbot for Customer Service on Social Media. In *Proceedings of CHI*, 3506–3510.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; et al. 2024a. Qwen2.5 Technical Report. *CoRR*, abs/2412.15115.

Yang, B.; Liu, D.; Xiao, C.; Zhao, K.; Tang, C.; Li, C.; Yuan, L.; Yang, G.; Huang, L.; and Lin, C. 2024b. Crafting Customisable Characters with LLMs: Introducing SimsChat, a Persona-Driven Role-Playing Agent Framework. *CoRR*, abs/2406.17962.

Ye, J.; Xiang, L.; Zhang, Y.; and Zong, C. 2025. SweetieChat: A Strategy-Enhanced Role-playing Framework for Diverse Scenarios Handling Emotional Support Agent. In *Proceedings of COLING*, 4646–4669.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020a. BERTScore: Evaluating Text Generation with BERT. In *Proceedings of ICLR*.

Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.-C.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; and Dolan, B. 2020b. DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation. In *Proceedings of ACL: System Demonstrations*, 270–278.